Quantification and visualization of carotid segmentation accuracy and precision using a 2D standardized carotid map

Bernard Chiu¹, Eranga Ukwatta²,³, Shadi Shavakh² and Aaron Fenster²,³

¹ Department of Electronic Engineering, City University of Hong Kong, Hong Kong
² Imaging Research Laboratories, Robarts Research Institute, London, Ontario, Canada
³ Graduate Program in Biomedical Engineering, University of Western Ontario, London, Ontario, Canada

E-mail: bcychiu@cityu.edu.hk, eukwatta@imaging.robarts.ca, sshavakh@alumni.uwo.ca and afenster@robarts.ca

Received 2 November 2012, in final form 12 March 2013
Published 8 May 2013
Online at stacks.iop.org/PMB/58/3671

Abstract

This paper describes a framework for vascular image segmentation evaluation. Since the size of vessel wall and plaque burden is defined by the lumen and wall boundaries in vascular segmentation, these two boundaries should be considered as a pair in statistical evaluation of a segmentation algorithm. This work proposed statistical metrics to evaluate the difference of local vessel wall thickness (VWT) produced by manual and algorithm-based semi-automatic segmentation methods (ΔT) with the local segmentation standard deviation of the wall and lumen boundaries considered. ΔT was further approximately decomposed into the local wall and lumen boundary differences (ΔW and ΔL respectively) in order to provide information regarding which of the wall and lumen segmentation errors contribute more to the VWT difference. In this study, the lumen and wall boundaries in 3D carotid ultrasound images acquired for 21 subjects were each segmented five times manually and by a level-set segmentation algorithm. The (absolute) difference measures (i.e., ΔT, ΔW, ΔL and their absolute values) and the pooled local standard deviation of manually and algorithmically segmented wall and lumen boundaries were computed for each subject and represented in a 2D standardized map. The local accuracy and variability of the segmentation algorithm at each point can be quantified by the average of these metrics for the whole group of subjects and visualized on the 2D standardized map. Based on the results shown on the 2D standardized map, a variety of strategies, such as adding anchor points and adjusting weights of different forces in the algorithm, can be introduced to improve the accuracy and variability of the algorithm.

(Some figures may appear in colour only in the online journal)
1. Introduction

Stroke is among the leading causes of death and disability worldwide, with estimated direct and indirect costs in 2010 of $73.7 billions in the United States alone (Lloyd-Jones et al 2010). Carotid atherosclerosis is a major source of thrombosis and subsequent emboli, which travel downstream and may block one of the cerebral arteries, causing ischemic stroke. Most strokes associated with carotid atherosclerosis can be prevented by lifestyle/dietary changes and medical/surgical treatments (Gorelick 1994, Spence 2007). Thus, improved identification of patients who are at risk for stroke as well as sensitive techniques for monitoring the response to therapy will have an enormous impact on the management of patients exposed to stroke risk.

Recent developments of 3D ultrasound imaging techniques have allowed detailed examinations of the anatomical structure of the carotid artery (Fenster et al 2004). Two major 3D biomarkers have been introduced for quantifying carotid atherosclerosis in 3D ultrasound, which are total plaque volume (TPV) (Landry et al 2004) and vessel wall volume (VWV) (Egger et al 2007, Krasinski et al 2009). Egger et al (2007) compared these two biomarkers and reported that VWV measurements were associated with higher intra-observer and inter-scan reproducibility than TPV. To obtain VWV measurements, segmentation of the lumen and outer wall boundaries are required. Although these boundaries can be manually segmented as in previous investigations and clinical studies (Chiu et al 2008b, 2008a, Egger et al 2008, Krasinski et al 2009), manual segmentation is laborious and time-consuming. Many segmentation algorithms have been developed to segment the lumen and/or outer wall on ultrasound (Gill et al 2000, Zahalka and Fenster 2001, Mao et al 2000, Ukwatta et al 2011), MRI (Yuan et al 1998, Zhang et al 2001, Luo et al 2003, Adame et al 2004) and CT angiography (Vukadinovic et al 2010). Segmentation results generated by these algorithms must be extensively validated before being used in clinical trials and subsequently in clinical settings. Many methods and metrics have been introduced for segmentation evaluation. Chalana and Kim (1997) described a contour comparison method in which they first generated a gold-standard boundary by averaging the manual segmentations performed by several observers. Contours generated by the algorithm were then matched with this gold-standard boundary in order to generate a point-by-point distance between the two boundaries. Ladak et al (2000), Mao et al (2000) and Chiu et al (2004) used similar methods to match the semi-automatically and manually segmented boundaries and summarized the point-by-point distances by the mean, maximum, mean absolute and root-mean-square distances. Chiu et al (2004) described area-based metrics to compute the area overlap and area difference between the semi-automatically and manually segmented boundaries. In addition to metrics used to evaluate segmentation accuracy, segmentation variability has also been assessed in the framework proposed by Gerig et al (2001), in which the intra- and inter-rater reliability of segmented volumes were assessed by the intra-class correlation coefficient (ICC). Local or point-by-point segmentation variability was considered in Mao et al (2000) and Gill et al (2000). Mao et al (2000) and Gill et al (2000) proposed 2D and 3D segmentation methods respectively to segment the carotid lumen boundaries and evaluated the local segmentation standard deviation of their algorithms. However, these metrics are only capable of evaluating anatomical structures that can be described by a single closed contour or surface.

The outer wall and lumen together define the size of the vessel wall, which is important for patient monitoring and clinical studies. Therefore, a pair of boundaries should be compared with another pair of boundaries in local segmentation evaluation of vascular segmentation algorithms. However, many investigations only evaluated lumen and wall segmentation separately (Klingensmith et al 2000, Adame et al 2004), providing limited information
regarding the accuracy and precision of the algorithm in measuring the size of the vessel wall. Since developing segmentation evaluation methods just amounts to defining metrics for comparing boundaries, in the following discussion, we do not restrict ourselves to previously proposed segmentation evaluation methods, but include methods of comparing vascular boundaries obtained in different settings, such as in inter-scan study (Egger et al 2007) or longitudinal study (Chiu et al 2008b). For comparing a pair of inner and outer boundaries with another pair, the most common metrics reported were vessel wall area (VWA) (Yuan et al 1998, Zhang et al 2001, Luo et al 2003, Chiu et al 2011) in selected transverse slices (e.g., the slice with maximum wall area in Yuan et al (1998) and Luo et al (2003)) and volume (VWV) (Egger et al 2007, Ukwatta et al 2011). In most of the VWA studies cited above, each scan was segmented once and therefore the segmentation variability was not studied, except in Yuan et al (1998), in which the slice 3 mm proximal to the bifurcation (BF) was measured twice by one observer and once more by another observer in order to assessed the intra- and inter-observer variability using ICC. Egger et al (2007) segmented each transverse slice five times for lumen and outer wall boundaries for 20 subjects, obtaining five VWV measurements for each subject. ICC and coefficient of variation were then computed to assess the intra-observer variability of VWV measurements.

Although VWV and VWA allowed simple comparisons between two pairs of boundaries, they provided limited information as to where the difference occurs within the artery. While VWV provided no spatial information regarding the vessel wall difference at all, VWA can be obtained in each transverse slice and therefore provide information for us to locate the slice where the difference occurred. However, the slice-by-slice VWA only provided a single-valued description for each slice and did not provide information about the locations where vessel wall difference occurred within a slice. In order to provide localized information on where the difference of the vessel wall occurred, our group has previously developed a vessel-wall-plus-plaque thickness (VWT) metric (Chiu et al 2008b). VWT is simply the distance between the outer and inner wall, which was computed on a point-by-point basis. The VWT-Change map was computed by first matching the VWT map generated at the two time points to obtain corresponding pairs and then subtracting the VWT value at follow-up by that at baseline. Although the VWT-Change map provided rich information on the distribution of VWT progression or regression, the VWT-Change maps between any two subjects cannot be compared, because the shape of the map depended on the geometry of the wall surface, which is highly subject-dependent. Although the flattened map approach we described (Chiu et al 2008a) facilitated a qualitative comparison between VWT-Change maps of different subjects, the flattened VWT-Change maps were still subject-dependent and quantitative local comparison was therefore not possible. To allow point-by-point comparison between different subjects or treatment groups, different carotid arteries must be mapped to a standardized 2D map. A method to generate such a map is introduced in this paper.

This paper describes a framework for segmentation evaluation that addresses the needs of (1) statistical comparison between a pair of boundaries (i.e., lumen and outer wall) segmented by an algorithm with the corresponding pair of manually segmented boundaries in vascular segmentation evaluation, (2) evaluation of local segmentation variability for the lumen and wall boundaries, (3) mapping the local segmentation accuracy and variability metrics to a standardized map for quantitative comparison. Local accuracy and variability metrics for the whole group of subjects involved in the evaluation can be mapped to this standardized map and averaged/pooled. The average difference or pooled standard deviation maps display a summary of the evaluation results obtained for a whole group of subjects, and are more representative in segmentation performance quantification than corresponding maps generated
for a single subject. This information allows for the identification of regions where the accuracy or variability of the algorithm requires to be improved.

2. Methods

2.1. Study subjects and ultrasound image acquisition

Since this study focuses on the evaluation of the segmentation results generated and evaluated in Ukwatta et al (2011), a brief summary of the subject and imaging acquisition techniques involved is provided here. 3D carotid ultrasound images of 21 subjects were segmented. All subjects provided written informed consent to the study protocol, which was approved by The University of Western Ontario Standing Board of Human Research Ethics. The image set was acquired from three subject groups: (1) Seven subjects had diabetes, (2) seven subjects suffered from rheumatoid arthritis and (3) seven subjects received atorvastatin treatment and with carotid stenosis exceeding or equal to 60%.

The 3D ultrasound images were acquired by translating an ultrasound transducer (L12-5, Philips, Bothel, WA) with a mechanical assembly along the neck of the subject for approximately 4.0 cm. The 2D ultrasound frames from the ultrasound machine (ATL HDI 5000, Philips, Bothel, WA) were digitalized at 30 Hz, and reconstructed into a 3D image (Fenster et al 2001, Landry et al 2005). The voxel size of the 3D ultrasound images were approximately $0.1 \times 0.1 \times 0.15 \text{ mm}^3$.

2.2. Image segmentation

This study focuses on analyzing and comparing manual and algorithm-based semi-automated segmentation methods based on five repeated segmentations produced by each method. These two segmentation methods are briefly summarized below.

2.2.1. Manual segmentation. The manual segmentation method was previously described and has been applied in Egger et al (2007) and Chiu et al (2008b). The reconstructed 3D ultrasound image was loaded into an in-house image analysis software and displayed using a multiplanar texture mapping approach (Fenster et al 2001). An expert observer first located the BF, and identified a medial axis parallel to the longitudinal axis of the common carotid artery (CCA). The 3D ultrasound images were resliced at 1 mm intervals by transverse planes that were perpendicular to the medial axis. In this study, 11 transverse 2D slices proximal to the BF were obtained, and the wall and lumen boundaries were manually segmented on each 2D slice. This process was repeated five times with a 24 h interval between consecutive segmentation sessions with the 21 3D ultrasound images randomized in each session. Thus, the expert observer was blinded to the image order, which would reduce memory bias.

2.2.2. Algorithm-based semi-automated segmentation. The same expert observer also selected eight points on the wall and lumen boundaries, of which four points were chosen to generate an initial boundary for the segmentation algorithm described in Ukwatta et al (2011). These manually identified boundary points were selected in the same image planes as in manual segmentation in order to allow direct comparison between segmentation results. Each initial boundary was then deformed using a level-set framework with local and global image forces described in Ukwatta et al (2011).
2.3. Mean and standard error of carotid thickness measurements

With five segmented contours obtained for each of the wall and lumen boundaries, we computed the mean and the standard error of carotid thickness on a point-by-point basis. This technique was introduced in a previous study (Chiu et al 2008b) and is summarized here for completeness.

The mean contours for repeated wall and the lumen segmentations were calculated separately using the following method. In computing the mean of five curves, these five curves were first matched pairwise using the modified symmetric correspondence method (Chiu et al 2008b). One contour (out of five segmented contours) with the largest circularity ratio, denoted as $C_1$, was chosen. Then $C_1$ was matched pairwise with the remaining four contours using the symmetric correspondence method. After the matching, each point on $C_1$ is associated with its corresponding points on the remaining four contours, forming a group of five corresponding points, denoted by $\{p_i: i = 1, 2, \ldots, 5\}$. The averages of these five-point groups were connected to form the mean curve. Figure 1(a) shows the mean curves (red curves) generated for repeated wall segmentations (black curves) and lumen segmentations (white curves).

To compute the mean thickness, $T$, at a point on the carotid wall, the symmetric correspondence algorithm was used to match the mean wall and the mean lumen boundaries obtained using the method described in the last paragraph. $T$ was defined to be the distance between each pair of corresponding points on the mean wall and lumen boundaries. Each pair of corresponding points, denoted by $p_W$ and $p_L$ in figure 1(a), defined a line that intersects the five wall and lumen boundaries. The distance between each of the five intersections on wall boundaries and $p_W$ can be obtained. The standard deviation of this group of five distances was defined as the local standard deviation of wall boundaries, denoted by $s_{W}$. Similar method can be used to define the local standard deviation of the lumen boundaries, denoted by $s_{L}$. $T$ can be expressed as $W - L$, where $W$ and $L$ represent the mean position of the wall and lumen with respect to a common origin and have standard errors $s_W/\sqrt{n_W}$ and $s_L/\sqrt{n_L}$ respectively, $n_W$ and $n_L$ represent the number of wall and lumen contours, respectively, that have an intersection with the line defined by the two points $p_W$ and $p_L$. In most cases, this line cuts all the contours and thus $n_W = n_L = 5$. In exceptional circumstance, $n_W$ or $n_L$ can be lower than 5. One such case is shown in figure 1(b), in which the five yellow corresponding lines could not intersect one of the lumen boundaries, giving $n_L = 4$. If this situation occurred, the contour without intersection would not be involved in the local standard deviation calculation. For the arteries investigated in this paper, this situation never happened for the wall boundaries (i.e., $n_W = 5$ for all arteries segmented either manually or algorithmically). It happened rarely for the lumen boundaries. For manual segmentation, $n_L < 5$ at 13 out of 9240 points investigated (40 points per slice per artery $\times$ 11 slices per artery $\times$ 21 arteries). For algorithm segmentation, $n_L < 5$ at 26 out of 9240 points. With $s_{W}$, $n_W$, $s_{L}$ and $n_L$, the standard error of $T$ can be computed:

$$SE_T = \sqrt{\frac{s_{W}^2}{n_W} + \frac{s_{L}^2}{n_L}}$$

with a degree of freedom (Satterthwaite 1946):

$$\nu_T = \frac{\left(\frac{s_{W}^2}{n_W} + \frac{s_{L}^2}{n_L}\right)^2}{\frac{s_{W}^2}{n_W} - 1 + \frac{s_{L}^2}{n_L} - 1}$$

In order to be able to perform a t-test to compare the mean thickness for algorithm segmentation, $\bar{T}_A$, and that for manual segmentation, $\bar{T}_M$, the standard error of $\Delta T = \bar{T}_A - \bar{T}_M$ and the associated degree of freedom, $\nu$, are required. The standard error is just the root of sum of squares of the respective standard errors of $\bar{T}_A$ and $\bar{T}_M$, and $\nu$ is computed using the method proposed by Satterthwaite (1946):
Figure 1. Method for computing the mean and standard deviations of VWT. (a) The black and the white contours respectively represent repeated segmentations of the wall and lumen boundaries. The two mean boundaries were matched on a point-by-point basis. These correspondence relationship defines lines that connect pairs of corresponding points, such as $p_L$ and $p_W$. These line intersects five wall and lumen boundaries. The mean VWT ($T$) was defined by the distance between $p_L$ and $p_W$. The local standard deviation can be computed by the positions of the intersections (black and white dots). (b) shows a special case in which one of the lumen boundaries (labeled lumen boundary 1) did not have intersection with five yellow corresponding lines. In this case, lumen boundary 1 was not involved in the calculation of the lumen segmentation standard deviation at these five points. $n_L$ denotes the number of lumen contours intersected by the corresponding lines. $n_L = 4$ instead of 5 at these five points.

$$SE_{\Delta T} = \sqrt{SE_{T,A}^2 + SE_{T,M}^2} \tag{3}$$
\[
\nu_{\Delta T} = \frac{(SE_{T,A}^2 + SE_{T,M}^2)^2}{(SE_{T,A}^2)^2 + (SE_{T,M}^2)^2}
\]

where variables with the second subscript \(A\) denote values for the algorithm segmentation, and those with the second subscript \(M\) denote values for the manual segmentation.

2.4. Mean and standard error of distances between corresponding boundaries generated using two methods

The difference in local thickness between two segmentation methods consists of two components: the distance between the mean wall boundaries and that between the mean lumen boundaries. The knowledge on how these two components contributed to the difference in local wall thickness would provide insights regarding where the segmentation error occurred, and improvement of the algorithm can be made.

The difference and the standard deviation between the algorithmically and manually generated wall boundaries were obtained similarly as for the mean and standard deviation of wall thickness described in the previous section. Instead of matching the wall and lumen boundaries, now the wall boundaries obtained using the two methods were matched. The mean wall boundary for the algorithm segmentation and the mean wall boundary for manual segmentation were matched by the modified symmetric correspondence algorithm in order to obtain a point-by-point signed distance. Figure 2(a) shows the result of this calculation for an example slice. The outer contours are the wall contours, with the colored contour representing the mean wall boundary for algorithm segmentation and the black contour representing the mean wall boundary for manual segmentation. The signed wall distance, \(\Delta W\), was superimposed on the mean contour for algorithm segmentation. The signed distance was defined to be positive if the mean contour for the algorithm segmentation is outside that for manual segmentation; otherwise, the signed distance is negative. The standard deviation of wall segmentation for the algorithm segmentation, \(s_{W,A}\), and manual segmentation, \(s_{W,M}\), are generated in the same manner as described in the previous section, by intersecting repeatedly segmented wall boundaries with the line connecting corresponding points between mean contours. The standard error of the mean distance between wall boundaries segmented by the two methods is equal to:

\[
SE_{\Delta W} = \sqrt{\frac{s_{W,A}^2}{n_{W,A}} + \frac{s_{W,M}^2}{n_{W,M}}}
\]

Point-by-point t-tests were performed in order to evaluate whether the mean distance is significantly different from 0. The \(t\) statistic and the degree of freedom are given by:

\[
t = \frac{\Delta W}{SE_{\Delta W}}
\]

\[
v = \frac{(s_{W,A}/n_{W,A} + s_{W,M}/n_{W,M})^2}{(s_{W,A}/n_{W,A})^2 + (s_{W,M}/n_{W,M})^2}
\]

\[
\nu = \frac{(s_{W,A}/n_{W,A})^2}{n_{W,A} - 1} + \frac{(s_{W,M}/n_{W,M})^2}{n_{W,M} - 1}
\]

where \(n_{W,A}\) and \(n_{W,M}\) are defined in the same way as in the previous section and equal to 5 if all wall contours can be intersected by the line connecting the two mean boundaries. The signed distance between the lumen boundaries associated with the two segmentation methods was obtained separately in exactly the same way. The result is also shown in figure 2(a).
Figure 2. Signed wall difference $\Delta W$ and lumen difference $\Delta L$ and the relationship with VWT difference $\Delta T$. (a) shows the semi-automatically segmented wall and lumen boundaries for an example slice with respectively $\Delta W$ and $\Delta L$ superimposed. The black contours represent the manually segmented wall and lumen boundaries. If the corresponding points among four mean contours were collinear, $\Delta T = T_A - T_M = \Delta W - \Delta L$. (b) shows a case where the algorithm (white) and manual (black) segmentation matches well, the correspondence points on four mean contours are close to collinear, therefore $\Delta T$ is well approximated by $\Delta W - \Delta L$. (c) shows a case where the algorithm undersegmented the lumen. The correspondence points are still close to collinear in most location except on the right hand side. The correlation between $\Delta W - \Delta L$ and $\Delta T$ for all subjects is plotted in figure 5.

According to figure 2(a), if the corresponding points among four mean contours: the mean wall and lumen boundaries associated with the two methods are collinear, the mean difference of wall thickness measurements using the two methods, $\Delta T = T_A - T_M$, can be
exactly decomposed into two components and expressed as $\Delta W - \Delta L$, which is the signed wall distance less the signed lumen distance between the two segmentation algorithms. However, thickness measurements for the algorithm and manual segmentation described in section 2.3 were performed separately. Therefore it is not possible to assume the corresponding points are collinear. It is also not reasonable to force these corresponding points to be collinear by intersecting all contours using the line connecting the corresponding points of the mean wall and the lumen boundaries of the algorithm (black lines in figure 2(c)). However, there were locations in which an intersection does not exist between the black line and the mean lumen boundary for manual segmentation (see black arrows in figure 2(c)). For these points, the mean thickness for manual segmentation would not even be defined. Nonetheless, we observed that in most slices evaluated, especially for cases such as that shown in slices figure 2(b) in which algorithm (white contours) and manual segmentation (black contours) matches well, the correspondence points on the four mean boundaries are close to collinear, and therefore the thickness difference would be well approximated by $\Delta W - \Delta L$. Even though for cases such as that shown in figure 2(c) in which the algorithm under-segmented the lumen, the correspondence points on the four mean boundaries are still close to collinear in most locations except on the right hand side. Thus, $\Delta T$ was still expected to be well approximated by $\Delta W - \Delta L$. In section 3.1, a correlation study will be performed to evaluate this hypothesis. The result would allow us to keep the interpretation of $\Delta W$ and $\Delta L$ in relation to the thickness difference $T_A - T_M$ in perspective.

2.5. Determining $\Delta T$, $\Delta W$ and $\Delta L$ based on median of repeated segmentations

Mean of repeated segmentations of wall and lumen boundaries were matched to compute $\Delta T$, $\Delta W$ and $\Delta L$ in sections 2.3 and 2.4. However, mean is not robust to outliers. In an ultrasound image, there may be a number of prominent edges that could be identified as the boundary and there were cases when the observer or the algorithm segmented different edges in different segmentation sessions. Figure 16 shows some examples where this situation occurred in lumen segmentation. For these cases, matching the median of boundaries to compute $\Delta T$, $\Delta W$ and $\Delta L$ may be more suitable since the median is less sensitive to outliers. In section 3.7, $\Delta T$, $\Delta W$ and $\Delta L$ will be computed based on the median of wall and lumen boundaries. The $\Delta T$, $\Delta W$ and $\Delta L$ maps thus generated will be compared with the corresponding maps generated using the mean boundaries.

2.6. Effect on distance from BF on segmentation accuracy and variability

Ukwatta et al (2011) evaluated the accuracy of the algorithm on a slice-by-slice basis separately for the wall and lumen boundaries and concluded that the accuracy of the algorithm is lower and the variability higher near the BF. We will evaluate this conclusion using the local evaluation metrics described in the above two sections. Each mean wall contour segmented by the algorithm for each subject was uniformly resampled at 40 points. Each point was equipped with three accuracy metrics, $\Delta T$, $\Delta W$ and $\Delta L$ as will be described in section 2.7.1. Therefore, for each slice, there are 880 data points (40 per subject per slice $\times$ 21 subjects) for each of these measurements. For each slice, the mean of each of these three measurements was computed to assess whether the algorithm over-segmented or under-segmented the wall and lumen in relation to manual segmentation, and whether the segmentation error resulted in bias for thickness measurements obtained using the segmentation algorithm. The means of the absolute values of the three local measurements, $|\Delta T|$, $|\Delta W|$ and $|\Delta L|$, were also computed to
assess the absolute segmentation error of the algorithm. One-way ANOVA was applied to test the null hypothesis that these six measurements associated with each slice were equal. If the null hypothesis was rejected, the Tukey test would be used for pairwise comparison between each measurement associated with each unique pair of slices.

Each sample point is also associated with four standard deviations, which are the standard deviations of each of the wall and lumen boundaries for each of the two segmentation methods (i.e., $sW_A$, $sW_M$, $sL_A$ and $sL_M$). For slice $i$, a pooled standard deviation, $SD(i)$, was computed separately for each of the four standard deviations as below:

$$SD(i) = \sqrt{\frac{\sum_{j=1}^{S} \sum_{k=1}^{K} SD^2(i, j, k)}{S \times K}}$$

where $SD(i, j, k)$ denotes any of the four standard deviations at point $k$ of slice $i$ of the artery of subject $j$, $S$ denotes the total number of subjects and $K$ denotes the number of sample points per subject per slice. $S = 21$ and $K = 40$ in this study. The degree of freedom of $SD(i)$ is $SK\nu$ where $\nu$ is the degree of freedom of $SD(i, j, k)$. As described in section 2.3, except for a very few cases where not all five repeated segmentations have an intersection with the line joining corresponding points, the number of samples used to calculate $sW_A$, $sW_M$, $sL_A$ and $sL_M$ is 5. Thus, these standard deviations have a degree of freedom of 4. Since the degrees of freedom are very large ($21 \times 40 \times 4 = 3360$), any difference between the standard deviations associated with two slices is statistically significant and the F-test was not performed.

2.7. Evaluation of segmentation methods for a whole group of subjects

In order to reach a conclusion applicable to the whole group of subjects, the carotid arteries of all subjects must be mapped to a standard shape in which the local evaluation metrics for different subjects can be compared and analyzed. We developed a standardized 2D rectangular map for this purpose. However, before describing the flattening map methodology, it is necessary to map all local evaluation metrics produced for each subject described in sections 2.3 and 2.4 onto a single surface to be flattened.

2.7.1. 3D local evaluation metrics map. The local evaluation metrics introduced thus far were all mapped to the mean wall surface for the algorithm segmentation, which we refer to as 3D local evaluation metrics map hereafter. At the end of the mapping operations described below, each point on the 3D local evaluation metrics map is equipped with the following list of local metrics:

- $T_A$: local mean thickness for algorithm segmentation.
- $SE_{T_A}$: standard error of $T_A$.
- $T_M$: local mean thickness for manual segmentation.
- $SE_{T_M}$: standard error of $T_M$.
- $\Delta T$: $|\Delta T|$, $T_A - T_M$ and its absolute value.
- $T_{\Delta T}$: results of point-by-point t-tests for $\Delta T$ (1 if significant, 0 otherwise).
- $\Delta W$, $|\Delta W|$: mean (absolute) distance between wall boundaries generated manually and by the algorithm.
- $T_{\Delta W}$: results of point-by-point t-tests for $\Delta W$.
- $\Delta L$, $|\Delta L|$: mean (absolute) distance between lumen boundaries generated manually and by the algorithm.
- $T_{\Delta L}$: results of point-by-point t-tests for $\Delta L$.
- $sW_A$: local standard deviation for wall boundaries generated by the algorithm.
- $sW_M$: local standard deviation for wall boundaries generated manually.
The following list summarizes the steps for generating the 3D local evaluation metrics maps:

(i) $T_A$ and $SE_{T,A}$ (equation (1)) were computed for each pair of corresponding points on the mean lumen and wall boundaries generated by the algorithm according to section 2.3. Both metrics were superimposed to the mean wall boundary.

(ii) $T_M$ and $SE_{T,M}$ were computed in the same way as above and superimposed to the mean wall boundary for manual segmentation, which was then matched to the mean wall boundary for algorithm segmentation using the symmetric correspondence method. $T_M$ and $SE_{T,M}$ of a point on the mean wall boundary for manual segmentation were mapped to its corresponding point on the mean wall boundary for algorithm segmentation.

(iii) Each point on the mean wall boundary for algorithm segmentation was now equipped with $T_A$, $T_M$, $SE_{T,A}$, $SE_{T,M}$. A point-by-point t-test was then performed according to equations (3) and (4), resulting in $T_{\Delta T}$, which was mapped on the mean wall boundary for algorithm segmentation.

(iv) $\Delta W$ was computed by matching the mean wall boundaries generated manually and by the algorithm. A point-by-point t-test was then performed according to equations (5)–(7), resulting in $T_{\Delta W}$. $\Delta W$ and $T_{\Delta W}$ were mapped to the mean wall boundary for the algorithm. $S_{W,A}$ and $S_{W,M}$ were computed by intersecting repeated segmentations with line joining corresponding points between the two mean boundaries, and mapped on the mean wall boundaries for the algorithm as well.

(v) Similarly, $\Delta L$, $T_{\Delta L}$, $S_{L,A}$ and $S_{L,M}$ were mapped to the mean lumen boundary for the algorithm. These values were then mapped to the mean wall boundary for the algorithm according to the correspondence relationship already established in step (i).

2.7.2. Generation of 2D standardized carotid maps. The 3D local evaluation metrics map was first transformed to a local coordinate frame, in which the external carotid artery (ECA) is located on the negative x side and the internal carotid artery (ICA) has positive x-coordinates. Although segmentation was only performed for the CCA for this study, at the CCA slice closest to the BF, the portion that would branch to ECA can be identified manually (figure 3(a)).

The directions of the local x-, y- and z-axes and a reference point that would map to the origin of the local frame must be defined for the transformation to the local coordinate axes. Segmentation was performed on parallel slices with inter-slice distance of 1 mm. Thus, it is natural to define the z-axis as the slice normal that points downstream. An expert observer segmented the part that would branch to ECA on the slice closest to the BF. This ECA contour intersects the mean lumen boundary at two points $p_1$ and $p_2$ as shown in figure 3(b). The midpoint between $p_1$ and $p_2$, denoted by BF, was defined to be the reference point that would map to the origin of the local axes. This point is analogous to the BF, although strictly speaking, the transverse slice still belongs to the CCA and has not reached the BF. The intersecting points $p_1$ and $p_2$ divided the mean lumen boundary into two segments. The segments on the left and right hand side of figure 3(b) have centroids $C_{ICA}$ and $C_{ECA}$ respectively. The x-axis was defined to be the vector pointing from $C_{ECA}$ to $C_{ICA}$. The y-axis was computed by taking the cross product of the z-axis and the x-axis. With these defined, the local coordinates $(x_L, y_L, z_L)$ of the 3D local evaluation metrics map can be obtained from its coordinates $(x_G, y_G, z_G)$ by the following transformation:
Figure 3. Generation of 2D standardized carotid maps. (a) The white solid lines represent the mean wall and lumen boundaries segmented by the algorithm. The dotted line represents the manually segmented external carotid (ECA) boundary. (b) shows the same boundaries as in (a) with the BF and the $x$-axis of the reference frame labeled. The intersections between the ECA boundary and the mean lumen boundary, labeled $p_1$ and $p_2$, divides the mean lumen boundary into two curves, which have centroids $C_{ECA}$ and $C_{ICA}$. (c) shows how each transverse slice of the 3D evaluation metrics map (section 2.7.1) was divided into two curves. $C_{Slice}$ is the point of intersection between the slice and the line $C_{UP}$-BF. (d) The 3D map of each artery was cut and divided into the ECA and ICA sides. The average surface areas of the two sides computed over the whole group of subjects were divided by the longitudinal coverage (10 mm in this study) to obtain $L_{ECA}$ and $L_{ICA}$ respectively. (e) shows the 2D standardized map. The dotted line and the green solid line in (c) were respectively mapped to corresponding lines here.

$$
\begin{bmatrix}
  x_L \\
  y_L \\
  z_L \\
  1 
\end{bmatrix} = \begin{bmatrix}
  x_u & y_u & z_u & BF \\
  0 & 0 & 0 & 1 
\end{bmatrix}^{-1} \begin{bmatrix}
  x_G \\
  y_G \\
  z_G \\
  1 
\end{bmatrix} \tag{9}
$$
where $x_u$, $y_u$, and $z_u$ are the $3 \times 1$ unit vectors of the $x$-, $y$- and $z$-axes, and $\mathbf{BF}$ is the $3 \times 1$ positional vector representing the reference point BF.

For each transverse slice of the 3D local evaluation metrics map described in section 2.7.1, a point $C_{\text{Slice}}$ was defined as the intersection between the slice and the line connecting the centroid of the slice farthest away from the BF ($C_{\text{UP}}$ in figure 3(d)) and the reference point BF (black line in figure 3(d)). The line centered at $C_{\text{Slice}}$ with direction parallel to the $y$-axis cut the outer wall boundary on each transverse slice at two points, $I_{C1}$ and $I_{C2}$, as shown in figure 3(c). These two intersection points divided the wall boundary on each slice into two curves: the dotted curve on the ECA side and the solid curve on the ICA side. These two lines were mapped to a rectangular domain parameterized by $(u, v)$ as shown in figure 3(e). The dotted curve was mapped to a horizontal line with length $L_{\text{ECA}}$ and $v$ equal its 3D $z$-coordinate. The solid curve was mapped to a line located at the same vertical position, but with length $L_{\text{ICA}}$, as shown in figure 3(e). $I_{C1}$ was mapped to $u = 0$. Because the closed contour was cut at $I_{C2}$, this point was mapped to two locations, $u = -L_{\text{ECA}}$ and $u = L_{\text{ICA}}$, on the 2D map. The inter-slice distance of 1 mm was preserved in the 2D planar map, but the lengths of the curves were scaled so that all transverse slices have the same length (i.e., $L_{\text{ECA}} + L_{\text{ICA}}$) in the 2D map although the arc-lengths of slices vary. However, the 3D to 2D mapping preserves the average area computed over the whole group of subjects and this area-preservation property holds for each of the ECA and ICA sides. This is achieved by selecting the lengths $L_{\text{ECA}}$ and $L_{\text{ICA}}$ in the following way. Each 3D map was cut by a plane containing the line $C_{\text{UP}}$ and the $y$-axis, dividing the 3D map into the ECA and ICA sides, which are the left and the right side respectively as shown in figure 3(d). The surface areas of the ECA and ICA sides were obtained for each 3D map. The average surface areas of the two sides were computed for the 21 arteries investigated in this study. The average area on each side was divided by the longitudinal coverage (10 mm in this study) to obtain $L_{\text{ECA}}$ and $L_{\text{ICA}}$ respectively. In this way, the proposed 3D to 2D mapping preserved the average surface area over the studied population for each of the ECA and ICA sides. That is a much more global definition of area-preservation in contrast with our previously proposed area-preserving mapping technique (Chiu et al. 2008a), which preserved the area of each infinitesimal region for each artery. Figure 4 shows the 2D standardized map generated for one subject with $\Delta T$ color-coded and superimposed.

2.7.3. 2D mean metrics maps for the whole group of subjects. Same as the 3D local evaluation metrics map, each point on the 2D standardized map for each subject was equipped with the local evaluation metrics listed in section 2.7.1. At each point $(x, y)$ on the 2D standardized map, the mean of a local metric $M$, denoted by $\overline{M}(x, y)$, was defined as:
Figure 5. The correlation between $\Delta W - \Delta L$ and $\Delta T$.

$$M(x, y) = \frac{1}{S} \sum_{i=1}^{S} M_i(x, y)$$  \hspace{1cm} (10)

where $M_i(x, y)$ represents the value of the local metric $M$ at point $(x, y)$ of Subject $i$’s 2D standardized map and $S$ represents the total number of subjects, which is 21 in our study. The local metric $M$ in this equation can be one of the following metrics: $\Delta T$, $|\Delta T|$, $\Delta W$, $|\Delta W|$, $\Delta L$ or $|\Delta L|$. Thus, a total of six standardized maps were generated, each with the average of one of these local metrics superimposed.

As described in section 2.6, each point is associated with four standard deviations. In section 2.6, the pooled standard deviations were computed for each slice. Here the pooled standard deviations were computed for each point on the 2D standardized template. The pooled standard deviation of a local standard deviation metric, $SD(x, y)$, was computed by:

$$SD(x, y) = \sqrt{\frac{1}{S} \sum_{i=1}^{S} SD_i^2(x, y)}$$ \hspace{1cm} (11)

where $SD_i(x, y)$ is the value of the standard deviation metric at point $(x, y)$ of subject $i$’s 2D standardized map and $S$ represents the total number of subjects. Local pooled standard deviation was computed for $s_{W_A}$, $s_{W_M}$, $s_{L_A}$ and $s_{L_M}$. Thus, a total of four standardized maps were generated for these standard deviation metrics.

3. Results

3.1. The relationship among $\Delta T$, $\Delta W$ and $\Delta L$

As described at the end of section 2.4, the difference in thickness obtained using manual and algorithm segmentations, $\Delta T$, exactly equals to $\Delta W - \Delta L$ only if the four corresponding points on the mean wall and lumen boundaries associated with the two methods are collinear. In determining the thickness measurements, the mean wall and the lumen boundaries associated with each segmentation method were separately matched. Thus, the corresponding points are not collinear. In this section, we evaluated how well $\Delta T$ was approximated by $\Delta W - \Delta L$ using Pearson correlation analysis. Figure 5 shows the plot of $\Delta W - \Delta L$ against $\Delta T$. The Pearson correlation coefficient is 0.96. Thus, we can consider $\Delta W - \Delta L$ a very close approximation of $\Delta T$ in the following analysis.
Figure 6. (a) shows the plot for the mean values of $|\Delta T|$, $|\Delta W|$ and $|\Delta L|$ computed for each transverse slice. Slice 10 is closest, while slice 0 farthest away, from the BF. (b) shows the plot for the mean values of $\Delta T$, $\Delta W$ and $\Delta L$ computed for each transverse slice. In (a) and (b), the error bars are the 95% confidence interval of the means. The metrics in (a) indicate the absolute error of the algorithm and those in (b) indicate bias in the estimation of the boundary (i.e., over-estimation or under-estimation) by the algorithm.

3.2. Average $|\Delta T|$, $|\Delta W|$ and $|\Delta L|$ at each slice

Figure 6(a) shows the average $|\Delta T|$, $|\Delta W|$ and $|\Delta L|$ calculated for each slice with slice 10 being closest and slice 0 farthest from the BF. The error bars on each curve represents the
Table 1. Results of multiple comparison tests for six metrics. The number represents the slice number. Slice 10 is closest, while slice 0 farthest away, from the BF. The slice number in this table was ordered in the ascending order of the respective metrics. Lines were drawn to group subsets of slices with means that are not statistically significant in the 0.05 significance level.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Subsets for $\alpha = 0.05$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>\Delta T</td>
</tr>
<tr>
<td>$</td>
<td>\Delta W</td>
</tr>
<tr>
<td>$</td>
<td>\Delta L</td>
</tr>
<tr>
<td>$\Delta T$</td>
<td>87 794620 946205101 62051013</td>
</tr>
<tr>
<td>$\Delta W$</td>
<td>1089 76052 605234 052341</td>
</tr>
<tr>
<td>$\Delta L$</td>
<td>10 935 3508612 5086127 0861274</td>
</tr>
</tbody>
</table>

95% confidence interval for the means calculated at each slice. All three metrics suggested that segmentation error increased as we approach the BF. One-way ANOVA for all three metrics rejected the null hypotheses that the respective means were equal at all slices with $p < 0.001$. The results of the Tukey’s multiple comparison tests are shown in table 1. The numbers in the table represents the slice number, which was ordered in the ascending order of the respective metrics. Lines were drawn to group subsets of slices with means that are not statistically significant. For all three metrics, slices 6–10 were associated with larger error than slices farther away from the BF.

3.3. Average $\Delta T$, $\Delta W$ and $\Delta L$ at each slice

Figure 6(b) shows the average $\Delta T$, $\Delta W$ and $\Delta L$ calculated for each slice. All three metrics suggested that the segmentation bias was close to 0 for slices 0–6. Again, one-way ANOVA for all three metrics rejected the null hypotheses that the respective means were equal at all slices with $p < 0.001$. Starting from slice 7, the algorithm under-segmented the wall boundary, and the effect was most pronounced at slices 8–10. The under-segmentation was statistically significant as indicated by the result of the Tukey test shown in table 1. The under-segmentation of the lumen boundary were pronounced only in slices 9 and 10 and the effect was only statistically significant at slice 10 according to Tukey test. Since $\Delta T \approx \Delta W - \Delta L$, the under-estimation of thickness by the algorithm is most pronounced in slice 8, with the wall being significantly under-segmented while no bias was observed in the lumen segmentation. In slice 10, both the under-segmentation of the lumen and wall canceled each other, resulting in a bias that is close to 0 in the thickness estimation.

3.4. Pooled standard deviations of wall and lumen boundaries

Figure 7 shows the pooled standard deviations of wall and lumen boundaries for manual and algorithm segmentations. The plots for manual segmentation are shown in solid lines, while those for the algorithm segmentation are shown in dotted lines. The results for lumen boundaries are plotted in red, while those for wall boundaries are plotted in blue. Pooled standard deviations tend to be higher close to the BF, except for the manually segmented lumen for which the pooled standard deviations at the first several slices farthest away from the BF were equally high. The pooled standard deviations for the wall boundaries segmented by the algorithm are consistently greater than manually segmented boundaries at all transverse image slices. The pooled standard deviations for lumen boundaries for both manual and algorithm segmentations are high at slices 0–3 in figure 7. In section 3.6, this region of high
variability in lumen segmentation was more precisely located and analyzed using the 2D standardized map.

3.5. Local metrics for a single subject mapped on the standardized template

In this study, a variety of quantities introduced in section 2.7.1 were color-coded and superimposed on the 2D and 3D maps. Regarding color map selection, Borkin et al (2011) pointed out that a diverging color map leaded to fewer errors in identifying regions with low endothelial shear stress compared to the rainbow color map. Being a perceptually ordered color scheme, a diverging color scheme leads to easier and more accurate interpretation if the task is to identify one or both extremes in the color scale. The diverging color scheme is thus suitable and was used for coding the absolute differences (i.e., $|\Delta T|$, $|\Delta W|$ and $|\Delta L|$) and the pooled standard deviations ($s_{W,A}$, $s_{W,M}$, $s_{L,A}$ and $s_{L,M}$) in this study. On the other hand, for the signed $\Delta T$, $\Delta W$ and $\Delta L$, in addition to the requirement of identifying the two extremes (i.e., smallest negative and largest positive differences), an equally important task is to identify regions where the difference is close to 0. In this case, more colors are required and the rainbow color map, which encodes small negative difference as blue, no difference as green and large positive difference as red, provides more detail and was used for proper interpretation of the signed quantities.

Figure 8(a) shows the mean wall surface of repeated algorithm segmentation with $\Delta W$ color-coded and superimposed. The black mesh represents the mean wall surface of repeated manual segmentations. Figure 8(c) shows the corresponding standardized 2D map. Figure 8(b) shows an example of the transverse slice with the mean wall boundaries of algorithm and manual segmentation shown in solid and dotted line respectively. The asterisk here represents

![Figure 7](image.png)
Figure 8. Wall segmentation evaluation for a single subject. (a) shows the mean wall surface for algorithm segmentation with $\Delta W$ color-coded and superimposed. The black mesh represents the mean wall surface for manual segmentation. (c) shows the standardized map with $\Delta W$ superimposed. (b) shows an example of the transverse slice with mean wall boundaries of algorithm and manual segmentation shown in solid and dotted line respectively. The location of this slice is shown in (a) and (c). In (b), the asterisk indicates the ECA side of the artery and the short white line intersecting the boundaries indicates the location where the 3D map was cut and unfolded. Note that the ECA is on the left of this image, whereas the ECA is on the right of figure 3(a). This depends on whether the right (figure 3(a)) or left (figure (b)) carotid artery was imaged. The gray and white arrows respectively point to regions where the algorithm under-estimated and over-estimated the wall boundary. The corresponding locations of these two regions in the standardized map were also indicated in (c). (d) and (e) shows the standardized map with local segmentation variability of the algorithm ($s_{W,A}$) and manual segmentation ($s_{W,M}$) respectively. (f) shows the t-test evaluation results on whether the local $\Delta W$ is statistically significant (equations (6) and (7)) (white indicates statistically significance, while black indicates otherwise).
the ECA side of the artery and the short white line intersecting the boundaries indicates the location where the 3D map was cut and subsequently unfolded. The location of this slice was also labeled in figures 8(a) and (c). As described in section 2.7.2, the ECA side was mapped to left-hand side of the standardized 2D map as shown in figure 8(c). This map shows that the wall boundary was under-segmented on the ECA side and over-segmented on the near side of the transducer (indicated by the white arrow). The 2D map also displays the small under-estimation of the wall boundary by the algorithm on the ICA side of the artery.

Figures 8(d) and (e) show the local segmentation standard deviation associated with algorithm \(s_{W,A}\) and manual segmentation \(s_{W,M}\) respectively superimposed on the 2D maps. Comparison of the two 2D maps shows that the local segmentation standard deviation is higher for algorithm segmentation. Both maps show that the region on the ECA side near the BF has the highest local standard deviation. This region corresponds to the region pointed to by the gray arrow in figure 8(b), where the image gradient is low. The local signal difference in a carotid ultrasound image along the outward normal direction of the boundary was computed in one of our previous publications (Chiu et al 2010). The results regarding local standard deviations in this paper agrees with the observation we made in that publication: Boundaries perpendicular to the direction of ultrasound transmission are associated with higher local signal difference and a lower standard deviation, while edges parallel to the transmission direction produced lower signal difference and a higher standard deviation. Figure 8(f) shows the result of the point-by-point t-tests performed for \(\Delta W\) (equation (6)). White indicates statistical significant difference and black indicates otherwise.

Figure 9(a) shows the mean lumen surface generated by algorithm segmentation with \(\Delta L\) color-coded and superimposed. The black mesh represents the mean lumen surface associated with manual segmentation. Figure 9(c) shows the corresponding standardized 2D map. Figure 9(b) shows an example of the transverse slice with the mean wall boundaries of algorithm and manual segmentation shown in solid and dotted line respectively. The location of this slice in figures 9(a) and (c) was labeled. The 2D map shows that the lumen boundary was under-estimated by the algorithm on the ECA side, which was also apparent by comparing the colored and the black meshes in figure 9(a). This under-estimation is the result of the boundary separation-based constraint of the algorithm, which drove the lumen boundary to maintain a distance of at least 0.5 mm from the wall boundary. As the wall boundary was under-estimated in this region, so was the lumen boundary. In addition to this under-estimation, it can be observed that the algorithm over-estimated the lumen boundary on the top of the boundary pointed to by the arrow in figure 9(b). For slices close to the BF, an expert observer, when performing manual segmentation, would expect that the boundary consists of two circles or ellipses overlapping each other in the middle. Thus, if the edge was not well-defined in the image, the observer tended to segment a contour with a concave section at the bottom and on top of the boundary. However, the algorithm has a regularization energy that made the boundary smooth, limiting the ability for the segmentation algorithm to model irregular boundaries, thereby over-segmenting the wall and lumen boundaries (white arrows in figures 8(b) and 9(b)).

Figures 9(d) and (e) show the local segmentation standard deviation associated with algorithm \(s_{L,A}\) and manual segmentation \(s_{L,M}\) respectively. As expected, both maps show high standard deviation on the ECA side (i.e., left side of the 2D map (see figure 4)) as the edge is not well-defined here with the boundary parallel to the direction of ultrasound transmission. Similar to wall segmentation, local standard deviation was higher for the algorithm than manual segmentation. Figure 9(e) shows that significant \(\Delta L\) predominantly occurred on the ECA side where the lumen boundary was under-estimated by the algorithm.
Figure 9. Lumen segmentation evaluation for the same subject shown in figure 8. (a) shows the mean lumen surface for algorithm segmentation with $\Delta L$ color-coded and superimposed. The black mesh represents the mean lumen surface for manual segmentation. (c) shows the standardized map with $\Delta L$ superimposed. (b) shows an example of the transverse slice with mean lumen boundaries of algorithm and manual segmentation shown in solid and dotted line respectively. The location of this slice is shown in (a) and (c). In (b), the asterisk indicates the ECA side of the artery and the short white line intersecting the boundaries indicates the location where the 3D map was cut and unfolded. The gray and white arrows respectively point to regions where the algorithm under-estimated and over-estimated the lumen boundary. The corresponding locations of these two regions in the standardized map were also indicated in (c). (d) and (e) shows the standardized map with local segmentation variability of the algorithm ($s_{L,A}$) and manual segmentation ($s_{L,M}$) respectively. (f) shows the t-test evaluation results on whether the local $\Delta L$ is statistically significant (white indicates statistically significance, while black indicates otherwise).
Figure 10. (a) shows the VWT difference ($\Delta T$) between algorithm and manual segmentation for the same subject shown in figures 8 and 9. (b) and (c) show the standard error of VWT measurement for algorithm and manual segmentation (equation (1)). (d) shows the t-test evaluation results on whether $\Delta T$ is statistically significant (equations (3) and (4)).

Figure 10(a) shows the 2D map with $\Delta T$ superimposed. Since $\Delta T \approx \Delta W - \Delta L$, underestimation in wall segmentation was canceled by lumen boundary under-estimation in the $\Delta T$ metric. This was the case for part of the region on the ECA side pointed to by the gray arrow in figure 10(a) with both the wall and lumen boundaries under-estimated. By comparing the $\Delta W$ (figure 8(c)) and $\Delta L$ (figure 9(c)) maps, it can be observed that the significant over-estimation (as indicated by the t-test map in figure 10(d)) of thickness occurring on the slice closest to the BF was the result of the under-estimation of the lumen boundary (gray arrow in figure 9(c)). There was also a region with high over-estimation of $\Delta T$ on the left side of figure 10(a). However, due to the high standard error in thickness measurement using algorithm segmentation, this region was not well represented in the t-test map in figure 10(d).
In other words, the $t$ statistic associated with $\Delta T$, $t_{\Delta T} = \Delta T / \text{SE}_{\Delta T}$, was not high enough to be considered significant even though $\Delta T$ was high in this region. Instead, $t$-tests were more sensitive in regions on the right where standard errors were smaller even though $\Delta T$ was smaller.

### 3.6. Results on whole group of subjects using standardized template

Figure 11 shows the 2D standardized map with average $\Delta T$, $\Delta W$ and $\Delta L$ superimposed. In the slice-based analysis of $\Delta W$ in section 3.3, we found that the wall boundary was under-estimated by the algorithm in slices 8–10. The result agrees with the average $\Delta W$ map shown in figure 11(b). Some examples of wall boundary under-estimation near the BF are shown in figure 12. In these figures, the short white lines intersecting the boundaries indicate the location where the arterial surface was cut and unfolded. The dotted lines with arrows on both ends indicate segments where the algorithm under-estimated the wall boundary. The difference between algorithm and manual segmentation occurred due to the fact that when performing manual segmentation for slices close to the BF, the expert observer expected that the contours consist of two overlapping ellipses with a concave section at locations where the two ellipses intersect. In many cases, such as in figures 12(a)–(c), the algorithm stopped expanding at what supposed to be the concave region of the overlapping ellipses, and therefore
Figure 12. The transverse slice closest to the BF (i.e., slice 10) for five subjects. The short white line intersecting the boundaries indicates the location where the 3D map was cut and unfolded. In cases shown in (a)–(c), the elliptical boundaries stopped expanding at the concave region between the internal and external carotids, thus falling short of capturing the whole wall boundary in regions labeled by the dotted arrows. This explains the under-estimation of the wall boundaries at regions pointed to by the black arrows in figure 11(b). In cases shown in (d) and (e), the boundaries expanded beyond the concave region (red arrows in (d) and (e)), which explains the over-estimation of wall boundaries shown in the middle of figure 11(b) (white arrow).
3694 B Chiu et al

fell short of segmenting the full wall boundary. This explains the blue area on the right and light blue area on the left of the average $\Delta W$ map. In other cases, such as in figures 12(d) and (e), the algorithm expanded beyond the concave region, resulting in the over-estimation of wall boundary at the concave region, which corresponds to the red region in the middle of the average $\Delta W$ map pointed to by the white arrow (figure 11(b)).

As mentioned previously, there was a boundary separation-based energy driving the lumen boundary so that its distance from the wall boundary is at least 0.5 mm. Therefore, it is not unexpected that the lumen boundaries were under-estimated by the algorithm at similar regions as for the wall boundaries. However, the under-estimation for the lumen boundaries was greater on the left side (ECA side) than the right side (ICA side) of the standardized map, whereas the under-estimation for the wall boundaries was greater on the ICA side. There was also a region on the $\Delta L$ map (indicated by the arrows in figure 11(c)) where the lumen boundary was over-estimated. By going through the $\Delta L$ maps for individual subjects such as that shown in figure 9(c), we found that this over-estimation was largely due to the segmentation result for the subject shown in figure 13. The black line in figure 13 shows the location where the over-estimation occurred on the 2D standardized map. The corresponding transverse ultrasound image is also displayed. As observed in figure 13, it is apparent that the over-estimation of the boundary was due to the fact that the algorithm incorporated an echolucent plaque inside the lumen boundary.

Figure 14 shows the 2D standardized map with average $|\Delta T|$, $|\Delta W|$ and $|\Delta L|$ superimposed. These maps agree with the results shown in figure 6(a) (section 3.2) where the slice-based mean $|\Delta T|$, $|\Delta W|$ and $|\Delta L|$ were higher in slices close to the BF. The absolute difference also tends to occur where segmentation bias occurred. For example, $|\Delta T|$ was highest in the same regions where lumen under-estimation and over-estimation occurred.

Figures 15(a) and (b) show the pooled standard deviation of the wall boundaries segmented by the algorithm and manually respectively. The pooled standard deviations shown in these two maps have a similar distribution, but the standard deviations for the algorithm were higher than that for manual segmentation, which agrees with the results we had in the slice-based analysis shown in figure 7.

Figure 7 shows that the pooled local standard deviation was higher for manually segmented lumen boundaries than those segmented by the algorithm in the first four slices. By comparing the pooled standard deviation maps computed for lumen boundaries segmented by the algorithm (figure 15(c)) and those segmented manually (figure 15(d)), we can visualize where the difference occurred. This spatial information together with the 2D standardized map for individual subjects allowed us to find out which subject contributed to high standard deviation in each region. The high standard deviation shown on the lower left-hand side of figure 15(d) was mainly contributed by subject A. The standard deviation maps generated for Subject A are shown for manual and algorithm segmentation in figures 16(A-i) and (A-ii) respectively. In figure 16(A-i), the slice with high standard deviation was highlighted using the black line and shown on the right. A high standard deviation appeared in this slice because there was a weak bright region that the expert observer decided in one segmentation session that it was outside the lumen, but a part of the lumen in the remaining four sessions. However, this weak bright region did not have an edge strong enough to stop the evolution of the algorithm. Thus all five boundaries included the weak bright region as a part of the lumen. The high standard deviation region shown on the lower right-hand side of figure 15(d) was mainly contributed by subject B. The standard deviation maps generated for subject B are shown for manual and algorithm segmentation in figures 16(B-i) and (B-ii) respectively. Slices 0 and 1 in figures 16(B-i) and (B-ii) were highlighted to demonstrate the segmentation results. In slice 0 of both the algorithm and manual segmentation, there was one lumen boundary that was
Figure 13. Oversegmentation of lumen boundary in an example subject. (a) shows the 2D standardized map with $\Delta L$ color-coded and superimposed. (b) shows the region where the lumen boundary was greatly oversegmented by the algorithm (white arrow). This region is also indicated by the white arrow in (a).

Figure 14. 2D standardized map with average (a) $|\Delta T|$, (b) $|\Delta W|$ and (c) $|\Delta L|$ superimposed.

very different from the remaining four boundaries, resulting in high local standard deviation. For slice 1, only manual segmentation exhibited high local standard deviation since two of the five boundaries were different from the remaining boundaries. The high standard deviation region shown on the upper right-hand side of figure 15(d) was mainly contributed by subject C. The standard deviation maps generated for subject C are shown for manual and algorithm segmentation in figures 16(C-i) and (C-ii) respectively. Here, the bright regions in slices 8 and 9 have a higher intensity than in subjects A and B. Thus, it was even more difficult to determine whether the region is a part of the arterial wall or lumen. In slices 8 and 9, intra-observer disagreement existed on whether the bright region was inside the lumen. In addition, the edge of the bright region was also poorly defined, resulting in a high local segmentation standard deviation as shown in figure 16(C-i). The local standard deviation of the boundaries segmented by the algorithm was much lower in slice 9 than manually segmented contours,
Figure 15. Pooled standard deviation maps of (a) algorithmically segmented wall ($s_{WA}$), (b) manually segmented wall ($s_{WM}$), (c) algorithmically segmented lumen ($s_{LA}$) and (d) manually segmented lumen ($s_{LM}$) boundaries. The high local segmentation standard deviation in different regions was attributed to three subjects shown in figure 16.

because all the five repeated segmentation considered the bright region to be inside the lumen. The local standard deviation in slice 8 was higher than slice 9 for the algorithm segmentation because the bright region was outside one boundary. However, the standard deviation was still lower for the algorithm segmentation compared to manual segmentation in this slice.

3.7. $\Delta T$, $\Delta W$ and $\Delta L$ computed based on median boundaries

In cases demonstrated in figure 16 where different features were segmented in different segmentation sessions, computing $\Delta T$, $\Delta W$ and $\Delta L$ based on median wall and lumen boundaries is more suitable. Figure 17 shows the 2D standardized map with average $\Delta T$, $\Delta W$ and $\Delta L$ computed based on the median wall and lumen boundaries. Comparing these maps with figure 11 shows that $\Delta T$, $\Delta W$ and $\Delta L$ are, on average, insensitive to whether the mean or the median boundaries were used.

For subjects shown in figure 16, the $\Delta L$ maps generated based on the mean and median lumen boundaries were compared in figure 18. In figure 18, all transverse images have the same spatial orientations as its correspondence in figure 16. In all transverse slices in
Figure 16. (A-i)–(C-i) show the local standard deviation map for manually segmented lumen boundaries \((s_{L,M})\) of subjects A–C respectively. (A-ii)–(C-ii) show the local standard deviation map for lumen boundaries segmented by the algorithm \((s_{L,A})\) for subjects A–C respectively. Black line(s) in each figure show the location of the transverse slice(s) shown on the right. These transverse slice images have repeated lumen segmentation boundaries superimposed. The short white lines intersecting the boundaries superimposed on the ultrasound image indicate the position where the 3D map was cut and unfolded.

figure 18, the red and the green lines represent the mean of the manually and algorithmically segmented contours respectively, whereas the orange and the light blue lines represent the median of the manually and algorithmically segmented contours respectively. The transverse image in figure 16(A-i) shows an outlier for the manual lumen segmentation. Thus, the mean
Figure 17. 2D standardized map with average (a) $\Delta T$, (b) $\Delta W$ and (c) $\Delta L$ superimposed. Here $\Delta T$, $\Delta W$ and $\Delta L$ were computed using the median wall and lumen boundaries.

manual outline was inside the median manual outline as pointed to by the white arrow in figure 18(A-iii). Since the mean and median contours for the algorithm segmentations are almost the same, $\Delta L$ computed based on mean contours is greater than that computed using median contours in the region pointed to by the black arrows in figures 18(A-i) and (A-ii). Similar effect occurred in slice 1 of subject B as shown in figures 18(B-i), (B-ii) and (B-iv). The situation was different for subject C. In both figures 18(C-iii) and (C-iv), the median lumen contours were inside the corresponding mean lumen contours. $\Delta L$ computed based on the median contours was greater than that computed using mean contours. Figure 19 shows the 2D maps for subjects A–C with $s_{LA}$ plus $\Delta L^2$ computed using the median boundaries superimposed, which can be considered as a metric for the performance of the segmentation algorithm.

4. Discussion

The ultimate goal of carotid artery segmentation is to quantify the size of the plaque burden and vessel wall for longitudinal monitoring, which requires the segmentation for both the lumen and the wall boundaries. Therefore, a comprehensive evaluation of a segmentation algorithm should assess its ability to produce accurate and precise plaque burden and vessel wall measurements. This paper proposed statistical metrics to evaluate point-by-point vessel wall thickness (VWT) based on repeated segmentations of lumen and wall boundaries obtained manually and using the algorithm. The VWT difference ($\Delta T$) associated with the algorithm and manual segmentation was further decomposed into two components: the wall thickness difference ($\Delta W$) and the lumen thickness ($\Delta L$) obtained using two segmentation methods. The segmentation algorithm segments the lumen and wall boundaries separately, and it is important
Figure 18. (A-i)–(C-i) show the maps with $\Delta L$ superimposed for subjects A–C respectively computed based on the mean of repeated segmentations. (A-ii)–(C-ii) show the $\Delta L$ maps for subjects A–C respectively computed based on the median of repeated segmentations. Black line(s) in the 2D maps show the location of the transverse slice(s) shown on the right. In all the transverse slices, the red and the green lines represent the mean of the manually and algorithmically segmented contours respectively, whereas the orange and the light blue lines represent the median of the manually and algorithmically segmented contours respectively.

Another important aspect of an evaluation framework is the ability to summarize the accuracy and precision of segmentation results for the whole group of subjects involved in the test. In the evaluation of their segmentation algorithm, Ukwatta et al (2011) reported that the segmentation accuracy is lower and the variability higher near the BF. In order to evaluate this conclusion, we summarized the $\Delta T$, $\Delta W$ and $\Delta L$ obtained for the whole group of subjects by plotting the mean $|\Delta T|$, $|\Delta W|$ and $|\Delta L|$ for each of the 11 slices investigated. Figure 6(a) shows that all three mean metrics were high at slices closest to the BF. Tukey test further showed that these three metrics were higher at slices 7–10 (i.e., the four slices closest to the BF) comparing to the rest of the slices in a statistically significant way (table 1). We also

to know whether $\Delta W$ or $\Delta L$ contribute more to the local VWT error ($\Delta T$). The parameters used for lumen and wall segmentation can then be adjusted according to this information. In this study, the $\Delta W$ and $\Delta L$ maps shown in figures 8(c) and 9(c) provided an answer to the question on which of the wall ($\Delta W$) or lumen segmentation error ($\Delta L$) contribute more to $\Delta T$ shown in figure 10(a).
computed the pooled local standard deviation of the algorithmically and manually segmented wall and lumen boundaries for each transverse slice and plotted the results in figure 7. We found that the pooled standard deviation was higher near the BF for both the algorithmically and manually segmented wall boundaries, whereas the pooled standard deviation was high at slices 0–3 as well as in slices 7–10 for algorithmically and manually segmented lumen boundaries.

Although these metrics confirmed that the accuracy is lower and standard deviation higher for slices closest to the BF, they still do not provide information as to where the accuracy is low and standard deviation high within a slice. In order to summarize the local accuracy and variability metrics on a point-by-point basis, these metrics were first mapped to a 2D standardized map for each subject. The average segmentation accuracy and pooled standard deviation for the whole group of 21 subjects were then computed on this map on a point-by-point basis. The 2D standardized map facilitated the identification of important differences between the algorithmically and manually segmented boundaries. This comparison would have been much more laborious and prone to variability if the standardized map were three-dimensional, in which case an observer would be required to study the map in different angles using 3D visualization tools. The advantages of the 2D standardized map was demonstrated in section 3.6. By comparing the \( \Delta W \) map in figure 11(b) with that generated for individual subjects, we identified the subject for which under-estimation and over-estimation of the wall boundaries occurred and where it occurred (figure 12). These observations exposed the weakness of the algorithm in segmenting the wall and lumen boundaries at locations.
where the artery is about to bifurcate. The pooled standard deviation maps in figure 15 allowed easy comparison of the variation associated with the algorithmically and manually segmented boundaries, especially the lumen boundaries. Comparison between the maps in figures 15(c) and 15(d) with the corresponding maps generated for individual subjects revealed the locations where the variability of the lumen boundary was high as shown in figure 16. These maps provided detailed explanation as to why the pooled local standard deviation of lumen segmentation was higher for manual segmentation than algorithm segmentation in Slices 0–3 as shown in figure 7.

In the evaluation framework developed in this paper, \( \Delta T \), \( \Delta W \) and \( \Delta L \) were computed based on the mean wall and lumen boundaries derived from repeated segmentations. In cases where there are more than one prominent edge that could be identified as the boundary and different edges were segmented in different segmentation sessions, such as the repeated segmentations of the lumen boundaries shown in figure 16, the mean boundary is not a good representation of the repeated segmentations because it would not lie on any prominent edge, and sit on the average of these edges instead. In this paper, we considered the possibility of computing \( \Delta T \), \( \Delta W \) and \( \Delta L \) based on the median wall and lumen boundaries. However, section 3.7 shows that \( \Delta T \), \( \Delta W \) and \( \Delta L \) were not sensitive to whether the mean or the median boundaries were used, due to the fact the situation mentioned above was not common in the arteries studied in this paper.

Although the median boundary may be a better representation of the repeated segmentations in the sense described above, the major disadvantage of computing \( \Delta T \), \( \Delta W \) and \( \Delta L \) using the median boundaries is that the statistical tests described in sections 2.3 and 2.4 cannot be used anymore. New statistical tests are required to be developed based on the density function of the median (Rider 1960), which would require a full publication to establish. The statistical conclusion arrived using the median-based model may be similar to the mean-based model we developed here, which is highly likely to be the case for the population we investigated in this paper as \( \Delta T \), \( \Delta W \) and \( \Delta L \) calculated using the median boundaries were similar to those computed using the mean boundaries. Thus, it is always wise to check the difference between \( \Delta T \), \( \Delta W \) and \( \Delta L \) calculated using the median and mean boundaries. The result would allow us to determine whether it is worthwhile to develop the median-based model.

As mentioned in the introduction, a novelty of this assessment model is its ability on assessing the accuracy and precision of an algorithm in segmenting a pair of boundaries that define a structure with a certain thickness. However, the application of this assessment model is not limited to evaluating segmentation involving thickness of structures. Since a component of the assessment model evaluated the accuracy and precision for the wall and lumen boundaries individually (equations (5)–(7)), this component can be used to evaluate the accuracy and precision of boundaries segmented for anatomical structures that are described by a single closed contour, such as the prostate, liver, kidney and lung.

Another focus of this paper was on locating regions where the boundaries segmented by the algorithm have a larger discrepancy with manual segmentation and higher segmentation variability. The proposed evaluation framework is equally applicable for evaluating segmentation variability of manual segmentation. In our group, manual segmentation was performed by observers who took part in a six-week training in which the observers segmented a number of carotid images on five segmentation sessions. The segmentation variability was now assessed by the coefficient of variance and ICC of VWV measurements (Vidal et al. 2008). Despite the training, the lumen segmentation is still high for a few selected images, as observed in figure 16. The proposed local standard deviation assessment metric can used as a tool to help observers further reduce segmentation variability by providing feedback during
the training sessions regarding the location of the regions where the segmentation variability is high and needs to be improved.

5. Conclusion

In conclusion, we developed a comprehensive framework for assessing, displaying and summarizing the accuracy and precision of carotid boundaries segmented by different methods. Firstly, this framework (a) provided measurements of point-by-point thickness difference (i.e., $\Delta T$) between the wall thickness measured from boundaries segmented using the algorithm and that from manually segmented boundaries and; (b) decomposed $\Delta T$ approximately into $\Delta W$ and $\Delta L$, the mean point-by-point distances between the algorithmically and manually segmented wall and lumen boundaries respectively. This decomposition allowed us to assess whether the thickness difference was primarily contributed by the difference in the wall segmentation or lumen segmentation. Secondly, this framework provided local segmentation standard deviation for wall and lumen boundaries segmented manually and using the algorithm. Thirdly, 2D standardized map with local evaluation metrics superimposed (metrics were listed in section 2.7.1) was generated for each subject (figures 8–10). Thus, point-by-point comparison for different metrics between different subjects can be performed. Finally, since metrics associated with all subjects were mapped to the 2D maps with standardized dimensions, average metrics map can be computed and visualized (figures 11, 14 and 15). These average maps were valuable for identifying the patterns in which boundaries segmented by the algorithm differ from the manually segmented boundaries over the whole group of subjects being studied.

Acknowledgments

This paper is supported by City University of Hong Kong Strategic Research grant no. 7002762. EU acknowledges the support by the Natural Sciences and Engineering Research Council (NSERC) Canadian Graduate Scholarship. AF gratefully acknowledges the financial support provided by the Canadian Institute of Health Research. AF holds a Canada Research Chair in Biomedical Engineering, and acknowledges the support of the Canada Research Chair Program.

References

Gorelick P B 1994 Stroke prevention: an opportunity for efficient utilization of health care resources during the coming decade Stroke 25 220–4
Rider P R 1960 Variance of the median of small samples from several special populations J. Am. Stat. Assoc. 55 148–50
Satterthwaite F E 1946 An approximate distribution of estimates of variance components Biometrics Bull. 2 110–4